JAMA Network Open

Original Investigation | Health Informatics

Association of Clinician Diagnostic Performance With Machine Learning–Based Decision Support Systems
A Systematic Review

Baptiste Vasey, MMed; Stephan Ursprung, MMed; Benjamin Beddoe, BSc; Elliott H. Taylor, BSc; Neale Marlow, MBBS; Nicole Bilbro, MD; Peter Watkinson, MD; Peter McCulloch, MD

Abstract

IMPORTANCE An increasing number of machine learning (ML)–based clinical decision support systems (CDSSs) are described in the medical literature, but this research focuses almost entirely on comparing CDSS directly with clinicians (human vs computer). Little is known about the outcomes of these systems when used as adjuncts to human decision-making (human vs human with computer).

OBJECTIVES To conduct a systematic review to investigate the association between the interactive use of ML-based diagnostic CDSSs and clinician performance and to examine the extent of the CDSSs' human factors evaluation.

EVIDENCE REVIEW A search of MEDLINE, Embase, PsycINFO, and grey literature was conducted for the period between January 1, 2010, and May 31, 2019. Peer-reviewed studies published in English comparing human clinician performance with and without interactive use of an ML-based diagnostic CDSSs were included. All metrics used to assess human performance were considered as outcomes. The risk of bias was assessed using Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) and Risk of Bias in Non-Randomised Studies–Intervention (ROBINS-I). Narrative summaries were produced for the main outcomes. Given the heterogeneity of medical conditions, outcomes of interest, and evaluation metrics, no meta-analysis was performed.

FINDINGS A total of 8112 studies were initially retrieved and 5154 abstracts were screened; of these, 37 studies met the inclusion criteria. The median number of participating clinicians was 4 (interquartile range, 3-8). Of the 107 results that reported statistical significance, 54 (50%) were increased by the use of CDSSs, 4 (4%) were decreased, and 49 (46%) showed no change or an unclear change. In the subgroup of studies carried out in representative clinical settings, no association between the use of ML-based diagnostic CDSSs and improved clinician performance could be observed. Interobserver agreement was the commonly reported outcome whose change was the most strongly associated with CDSS use. Four studies (11%) reported on user feedback, and, in all but 1 case, clinicians decided to override at least some of the algorithms' recommendations. Twenty-eight studies (76%) were rated as having a high risk of bias in at least 1 of the 4 QUADAS-2 core domains, and 6 studies (16%) were considered to be at serious or critical risk of bias using ROBINS-I.

CONCLUSIONS AND RELEVANCE This systematic review found only sparse evidence that the use of ML-based CDSSs is associated with improved clinician diagnostic performance. Most studies had a low number of participants, were at high or unclear risk of bias, and showed little or no consideration for human factors. Caution should be exercised when estimating the current potential of ML to

(continued)

Key Points

Question Is clinician diagnostic performance associated with the use of machine learning–based clinical decision support systems?

Findings In this systematic review of 37 studies, no robust evidence was found to suggest an association between the use of machine learning–based clinical algorithms to support rather than replace human decision-making and improved clinician diagnostic performance.

Meaning Caution should be observed when estimating the current ability of machine learning algorithms to affect patient care, and emphasis on the evaluation of the human-computer interaction is needed.

+ Invited Commentary

+ Supplemental content

Author affiliations and article information are listed at the end of this article.

Abstract (continued)

improve human diagnostic performance, and more comprehensive evaluation should be conducted before deploying ML-based CDSSs in clinical settings. The results highlight the importance of considering supported human decisions as end points rather than merely the stand-alone CDSSs outputs.

JAMA Network Open. 2021;4(3):e211276. doi:10.1001/jamanetworkopen.2021.1276

Introduction

Artificial intelligence has been a popular term in the medical literature and health care industry for some time. Although we are still far from true artificial intelligence, advances in mathematical modeling and computing power have led to an increase in the number of published algorithms. Claims regarding the potential of artificial intelligence in medicine range from being of use to clinicians in their decision-making to artificial intelligence outperforming human experts. Funding for artificial intelligence in health care increases year after year,[1] and regulatory agencies are approving a growing number of software as medical devices (SaMDs) based on advanced machine learning (ML) algorithms, mainly in medical imaging.[2] Recent evidence suggests that the best-performing systems are now matching human experts' performance.[3] However, few randomized clinical trials or prospective studies have been carried out, and most nonrandomized trials in the field are at high risk of bias.[4]

Machine learning–based clinical decision support systems (CDSSs) are a category of SaMDs designed to support health professionals' decision-making by providing patient or problem-specific information learned from an ideally large number of clinical cases during a training process. Despite their name, most CDSSs are currently evaluated exclusively against human experts but rarely on their outcome when used with human clinicians of different seniority. Demonstrating that computers can be as good as humans for diagnostic tasks has some useful applications, notably for large population screening in which patients may otherwise not be able to see a physician in a timely manner. Nevertheless, this approach neglects an important factor in any medical encounter: the human clinician. As long as physicians hold the ultimate responsibility for signing off a diagnosis or treatment plan, it will be their interpretation of a CDSS output—not the output itself—that will affect patient care. Human decision-making is known to be influenced by numerous external factors and cognitive bias.[5-7] It would be unwise to assume without further evidence that a human operator would follow a diagnostic CDSS recommendation without question. Extending this argument further, we also have little evidence on how patients would react to fully automated diagnoses or treatment planning. It is therefore important to evaluate any new CDSSs in terms of its performance when used in interactive collaboration with a human clinician and not solely on its performance in silico (ie, on a test data set).

Previous systematic reviews have investigated the association of CDSSs with clinician performance or its surrogate clinical outcomes.[8-12] However, most of the included studies described systems whose parameters were defined by their developers or diagnosis generators based on handcrafted knowledge bases, hence not fully representing the true promise of ML: to become better than its creator by "learn[ing] without being explicitly programmed."[13] In this systematic review, we investigated the current evidence regarding the association between the use of ML-based diagnostic CDSSs and human performance and the ways these systems are evaluated by including all retrieved studies comparing human clinicians performing a diagnostic task with and without ML-based CDSS assistance.

Methods

Search Strategy and Selection Criteria

We conducted a systematic review of the literature, and this study followed the relevant sections of the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) reporting guideline.[14] The study is registered with PROSPERO (CRD42019140075).

A search strategy built around 4 additive concepts (machine learning, decision support system, clinician, and performance evaluation) was designed with the support of a specialist librarian and can be found in the study protocol (eAppendix 1 in the Supplement). The search was conducted in MEDLINE, Embase, and PsycINFO for the period between January 1, 2010, and May 31, 2019. The initial search was conducted on May 20, 2019, and the last search to identify possible late indexation within the specified time window was conducted on June 1, 2020. One round of a systematic forward and backward references search was conducted for all included studies. An additional search was performed using the names of algorithms recently approved by the US Food and Drug Administration. A grey literature search including the World Health Organization International Clinical Trials Registry Platform, conference abstracts (from 2017 onward), and the Cochrane Central Register of Controlled Trials was performed using an adapted search strategy (eAppendix 2 in the Supplement).

Inclusion criteria were peer-reviewed articles published in the English language, human physicians as the study population, the interactive use of an ML-based diagnostic CDSSs as intervention, human physicians without CDSSs as a control, any variable used to measure human performance as the main outcome, any variable to measure the stand-alone computer performance (ie, the performance achieved by the computer outputs without subsequent human intervention), and any variable describing the evaluation of the CDSS by the human operator as a secondary outcome. A CDSS was considered diagnostic if its output produced qualitative information (eg, benign vs malignant) about the nature of a lesion or if the detection of a lesion was in itself sufficient to pose a diagnosis and influence a therapeutic choice (eg, the presence of pulmonary emboli). Exclusion criteria were monitoring, alert, or detection-only systems; systems based on validated scores only; systems based on natural language processing only; and systems relying on handcrafted knowledge or rule bases. The specific definition of key concepts and complete exclusion criteria can be found in eAppendix 1 in the Supplement. All retrieved titles and abstracts were independently screened by at least 2 of us (B.V., S.U., E.H.T., N.M., and N.B.). Conflicts were adjudicated by a third reviewer (S.U. or B.B.). Full-text articles were independently reviewed for eligibility by at least 2 of us (B.V., S.U., B.B., E.H.T., N.M., and N.B.). Conflicts were resolved in consensus. The abstract screening and full-text review were conducted using the Covidence software.[15]

Data were extracted about the study population, patient population, data set characteristics, experiment description, system characteristics, metrics assessing human performance, metrics assessing computer performance, and study funding. The full list of data items can be found in eAppendix 1 in the Supplement. Investigators were not contacted. The risk of bias for each included study was assessed using the Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) tool as modified by Riches[8,16] and the Risk of Bias in Non-Randomised Studies–Intervention (ROBINS-I) tool.[17] QUADAS-2 was used to assess the risk of bias regarding the claims of CDSS diagnostic accuracy, and ROBINS-I was used to consider the risk of bias in the results assessing difference in performance. Studies were included in the analysis independently of their risk of bias. Data extraction and bias assessments were all conducted independently by at least 2 of us (B.V., S.U., B.B., E.H.T., N.M., and N.B.) using piloted forms. Conflicts were resolved by consensus. To ensure consistency, the main reviewer (B.V.) screened all abstracts and full texts for eligibility, extracted data, and assessed risk of bias on all included studies.

Meta-bias was investigated by searching the World Health Organization International Clinical Trials Registry Platform and Cochrane Central Register of Controlled Trials registers looking for

unpublished trials and evidence of selective reporting. The origin of study funding and the presence of a protocol were also considered.

Data Analysis

Narrative summaries were produced for the primary and secondary outcomes. As per protocol, subgroup analyses were performed for clinicians' experience level (experienced vs novice), the mathematical model used, the models' degree of support (single output vs information about process), and the reader paradigm (first vs second reader). First reader support displays the model output at the same time as the clinical data, and second reader support displays the model output after the observer had a chance to make their own decision on a case. An additional subgroup analysis for studies evaluating ML-based CDSSs in a representative clinical environment (clearly reported consecutive or nonaugmented random patient sample and access to the usually available clinical data at the time of decision-making) was performed. Patient-level results were prioritized over lesion-level results for the summary of main results. Patient or lesion types subgroup analyses were summarized separately. Given the heterogeneity of medical conditions, outcomes of interest, and evaluation metrics, no meta-analysis was performed. All studies were included in the analysis irrespective of their risk of bias.

Results

A total of 8112 titles were identified, of which 2774 were duplicates and 184 were not published in English; 5154 abstracts were screened, and 156 of these were selected for full-text review. Of the 156 studies assessed, 22 were eligible for inclusion. Fifteen additional publications meeting the inclusion criteria were retrieved from other sources, including forward/backward references search, trade names search, related literature references search, and publications tracing from the grey literature search. Thirty-seven publications were eventually included in this review.[18-54] **Figure 1** presents the PRISMA flowchart.

All included studies described CDSSs based on imaging modalities, with breast and pulmonary diseases being the most common medical conditions. Twenty studies (54%) investigated CDSSs technology with a designated trade name at the time of publication.

Thirty-one studies (84%) assessed a CDSS belonging to the International Medical Device Regulators Forum's risk category 4 (the highest category),[55] 25 studies (68%) used a second reader paradigm (see Data Analysis section) for the CDSS support, 8 (22%) used a first reader paradigm, 1 study (3%) used both, and 3 studies (8%) did not specify. Three studies (8%) used the same cases as training and test sets, and 3 studies (8%) did not report clearly on test set independence. The median proportion of events (ie, target condition) in the test set was 44% (interquartile range, 32%-54%). The median number of clinician participants in the observer test was 4 (interquartile range, 3-8), with each reading a median of 123 different cases (interquartile range, 79-300). **Table 1** gives an overview of the included studies' characteristics.

The 10 most common metrics used to quantify human performance in the included studies were sensitivity (81%), specificity (70%), area under the receiver operating curves (51%), accuracy (38%), interobserver agreement (30%), positive predictive value (PPV) (30%), negative predictive value (30%), reading time (22%), rate of recall for further investigation (11%), and the positive value of further investigations (8%). Equivalent metrics have been aggregated. A full list of evaluation metrics with their occurrence can be found in eTable 1 in the Supplement.

Table 2 reports a summary of the association between CDSS use and the 10 most common human performance evaluation metrics. Three studies reported on more than 1 CDSS (or using the same CDSS in different modalities).[21,31,42] A total of 107 main results were reported with statistical significance, and 41 were reported without it. Most studies defined statistical significance at $P < .05$, with some applying correction for multiple comparisons. Of the results reported with statistical significance, 54 studies (50%) showed an increase in their metrics, 4 (4%) reported a decrease, and

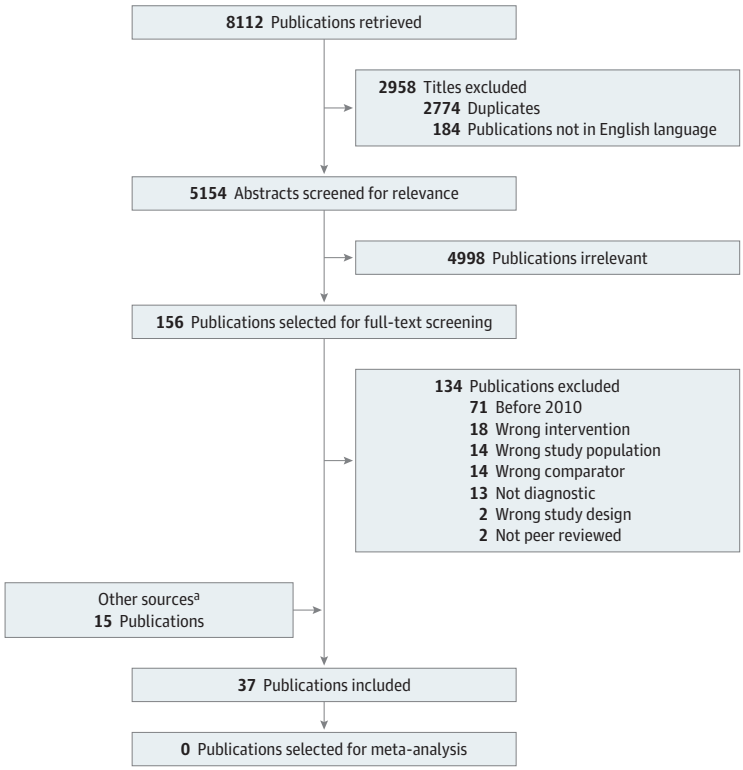
49 (46%) noted no change or an unclear change. The area under the receiver operating curves, accuracy, interobserver agreement, and PPV were usually increased with interobserver agreement showing the clearest change. The sensitivity, specificity, negative predictive value, rate of recall for further investigation, and PPV of further investigations remained unchanged in most cases, and the CDSS association with reading time showed no clear pattern. Sixteen studies also reported analyses on subgroups of patients or lesion types. A summary of these additional analyses can be found in eTable 2 in the Supplement, and a detailed list of the included studies' results are reported in eTable 3 in the Supplement.

In the 6 studies[20,22,24,29,39,41] evaluating CDSSs in a representative clinical environment for the same 10 evaluation metrics, 20 results were reported with statistical significance. Of these, 16 (80%) showed no difference in performance, and 4 (20%) reported an increase in sensitivity, area under the receiver operating curves, PPV, or interobserver agreement (eTable 4 in the Supplement).

In 19 studies in which a comparison was possible, CDSSs were more often associated with an increase in performance for less experienced clinicians compared with their senior colleagues (eTable 5 in the Supplement). The reader paradigm also appeared to be associated with human performance, with studies investigating CDSSs in the second reader mode appearing to be more often associated with an increase in the metrics (eTable 6 in the Supplement). The subgroup analyses according to the mathematical model used (eTable 7 in the Supplement) and degree of support (eTable 8 in the Supplement) produced no additional findings.

Twenty-seven studies reported on CDSS stand-alone performance. With the exception of 1 unclear case,[50] human participants always decided to override at least some of the CDSS recommendations. Of the 75 main results reported using the 10 most commonly applied metrics, the human contribution changed the system performance in 70 cases (93%). Compared with the stand-alone computer performance, adding human intelligence increased the metrics value in 45 cases (60%) and decreased it in 25 cases (33%). Only 3 results (4%) mentioned statistical significance; of

Figure 1. Flowchart of Study Inclusion



[a] Other sources included forward/backward literature search, reference search from relevant literature, trade name search, and conference abstracts or entries in the Cochrane Central Register of Controlled Trials that led to publications.

Table 1. Characteristics of Included Studies

Source	Medical condition	Algorithm used	IMDRF risk category	No. of sites	Test set sample size[a]	Test set event cases[b]	No. of study participants	Cases read/participant[c]	Reader paradigm	Private sector funding
Aissa et al,[18] 2018	Melanoma	ClearRead CT (CNN)[d]	4	1	46	46	3	46	First	NA
Aslantas et al,[19] 2016	Bone metastasis	Perceptron-based ANN	4	2	130	100	1	130	NA	No
Bargalló et al,[20] 2014	Breast cancer	SecondLook[d]	4	NA	21 321	130	4	8100[e]	Second	NA
Barinov et al,[21] 2019	Breast cancer	cCAD (ANN)[d]	4	Multiple	500	150	3	450 and 500	First and second	NA
Bartolotta et al,[22] 2018	Breast cancer	S-Detect (CNN)[d]	4	NA	300	122	4	300	Second	NA
Bien et al,[23] 2018	Knee musculoskeletal injury	CNN	3	2	120	99	9	120	First	NA
van den Biggelaar et al,[24] 2010	Breast cancer	SecondLook[d]	3	1	1048	50	2	524[e]	First	No
Blackmon et al,[25] 2011	Pulmonary embolism	VA10 PE (SVM)	4	NA	79	32	2	79	Second	NA
Cha et al,[26] 2019	Bladder cancer	CNN[d]	4	NA	123	40	12	123	Second	No
Chabi et al,[27] 2012	Breast cancer	B-CAD v.2[d]	4	1	160	77	4	160	Second	NA
Cho et al,[28] 2018	Breast cancer	S-Detect (CNN)[d]	4	1	119	54	2	119	Second	Yes
Choi JH et al,[29] 2018	Breast cancer	S-Detect (CNN)[d]	4	1	200	12	4	100[e]	Second	No
Choi JS et al,[30] 2019	Breast cancer	S-Detect (CNN)[d]	4	1	253	80	4	253	Second	No
Cole et al,[31] 2014	Breast cancer	ImageChecker v.1.0 (CNN) and SecondLook v.1.4[d]	4	Multiple	300 and 300	150 and 150	15 and 14	300 and 300	Second	No
Endo et al,[32] 2012	Pulmonary nodule	Euclidian distance clustering	4	1	30	23	3	30	NA	NA
Engelke et al,[33] 2010	Pulmonary embolism	PE-CAD (MIC)[d]	4	NA	58	58	4	58	Second	NA
Giannini et al,[34] 2017	Prostate cancer	SVM	4	NA	89	35	3	89	First	No
Hwang et al,[35] 2019	Thoracic pathology	CNN	4	1	200	103	15	200	Second	No
Lindsey et al,[36] 2018	Wrist fracture	CNN	3	1	300	NA	24	300	Second	Yes
Park et al,[37] 2019	Breast cancer	S-Detect (CNN)[d]	4	1	100	41	5	100	Second	No
Rodríguez-Ruiz et al,[38] 2019	Breast cancer	Transpara v.1.3.0 (CNN)[d]	4	2	240	100	14	240	NA	Yes
Romero et al,[39] 2011	Breast cancer	Image Checker v.5.4 (CNN)[d]	4	1	9389	124	2	4695[e]	Second	NA
Samulski et al,[40] 2010	Breast cancer	Image Checker v.8.0 (CNN)[d]	4	NA	120	40	9	120	Second	No
Sanchez Gómez et al,[41] 2011	Breast cancer	SecondLook v.1.1[d]	4	NA	21 855	94	6	3643[e]	Second	No
Sayres et al,[42] 2019	Diabetic retinopathy	CNN	3	Multiple	1796	213	10	1796	First	Yes
Shimauchi et al,[43] 2011	Breast cancer	Bayesian ANN	4	2	60	30	6	60	Second	NA
Sohns et al,[44] 2010	Breast cancer	Image Checker v.2.3 (CNN)[d]	4	NA	303	98	2	303	First	NA
Steiner et al,[45] 2018	Breast cancer	CNN	4	2	70	38	6	70	First	Yes
Stoffel et al,[46] 2018	Breast cancer	ViDi Suite v.2.0 (ANN)[d]	4	1	33	11	4	33	First	No
Sun et al,[47] 2014	Atrial thrombus	ANN	3	1	130	31	5	130	Second	No
Sunwoo et al,[48] 2017	Brain metastasis	k-means clustering + ANN	4	1	60	30	4	60	Second	No
Tang et al,[49] 2011	Ischemic stroke	ANN	4	Multiple	71	40	6	40	Second	No
Taylor et al,[50] 2018	Parkinsonian syndromes	SVM	3	1 and Multiple	55 and 100	33 and 60	2	55 and 100	Second	No
Vassallo et al,[51] 2019	Lung metastasis	SVM	4	1	225	75	3	225	Second	No
Watanabe et al,[52] 2019	Breast cancer	cmAssist (CNN)[d]	4	1	122	90	7	120	Second	Yes
Way et al,[53] 2010	Lung cancers	Linear discriminant analysis	4	1	256	124	6	NA	Second	No
Zhang et al,[54] 2016	Lymph node cancers	SVM	4	1	178	87	10	178	Second	No

Abbreviations: ANN, artificial neural network; CNN, convolutional neural network; IMDRF, International Medical Device Regulators Forum; MIC, multiple instance classifier; NA, not available; SVM, support vector machine.

[a] For the observer test with support for clinical decision support systems.

[b] One case can display multiple events.

[c] Each case was either seen once or multiple times (with or without assistance) depending on the study.

[d] Commercial name of proprietary algorithm.

[e] Mean value.

these, 1 showed no statistical difference and 2 noted a significant increase in accuracy. An overview of these results is reported in **Table 3**, a summary of the subgroup analyses in eTable 9 in the Supplement, and a detailed list of the results in eTable 10 in the Supplement.

Of the 37 included studies, 15 (41%) attempted to increase the interpretability of the model by presenting some of the intermediary calculations leading to the models' final output, and 13 studies (35%) included users' training before starting data collection. Four (11%) reported on user feedback about the CDSSs, of which 3 (8%) gathered feedback through a formalized process. Van den Biggelaar et al[24(p501)] asked study participants to indicate on their case evaluation forms if the CDSS marks "added valuable diagnostic information to their own original evaluation" but did not report on this outcome. Taylor et al[50(p5)] designed open and closed question interviews to "provide an insight into the CADx [computer-aided diagnosis]-radiologist relationship [and] to assess the effects of the CADx software on clinician decision-making." The study participants reported good agreement between their decision and the CDSS outputs, with a small to moderate influence on their reporting

Table 2. Association Between ML-Based CDSS Use and Clinician Performance[a]

Metric category	Results reported with statistical significance, No.			Results reported without statistical significance, No.			Total CDSSs evaluated, No.[b]
	Increase overall or for ≥50% of the participants	No change or unclear change for the group	Decrease overall or for ≥50% of the participants	Increase overall or for ≥50% of the participants	No change or unclear change for the group	Decrease overall or for ≥50% of the participants	
Sensitivity	10	11	1	9	1	0	32
Specificity	6	11	1	3	2	5	28
Area under the receiver operating curves	13	7	0	1	0	0	21
Accuracy	8	4	0	5	1	0	18
Interobserver agreement	7	2	0	2	0	0	11
Positive predictive value	5	3	0	2	0	2	12
Negative predictive value	3	5	0	3	1	0	12
Reading time	2	2	2	0	1	1	8
Recall for further investigations	0	2	0	1	0	0	3
Positive predictive value of further investigations	0	2	0	0	0	1	3

Abbreviations: CDSSs, clinical decision support systems; ML, machine learning.

[a] Number of main results reported for the 10 most commonly used metrics groups comparing computer-assisted clinicians with clinicians alone.

[b] Three studies reported on more than 1 CDSS or used the same CDSS in different modalities.

Table 3. Association Between Human Contribution and System Performance[a]

Metric category	Results reported with statistical significance, No.			Results reported without statistical significance, No.			Total CDSSs evaluated, No.[b]
	Increase overall or for ≥50% of the participants	No change or unclear change for the group	Decrease overall or for ≥50% of the participants	Increase overall or for ≥50% of the participants	No change or unclear change for the group	Decrease overall or for ≥50% of the participants	
Sensitivity	0	0	0	11	2	10	23
Specificity	0	0	0	11	2	5	18
Area under the receiver operating curves	0	1	0	5	0	7	13
Accuracy	2	0	0	6	0	1	9
Interobserver agreement	0	0	0	6	0	0	6
Positive predictive value	0	0	0	0	0	0	0
Negative predictive value	0	0	0	4	0	2	6
Reading time	0	0	0	0	0	0	0
Recall for further investigations	0	0	0	0	0	0	0
Positive predictive value of further investigations	0	0	0	0	0	0	0

Abbreviation: CDSSs, clinical decision support systems.

[a] Number of main results reported for the 10 most commonly used metrics groups comparing computer-assisted clinicians with stand-alone computers.

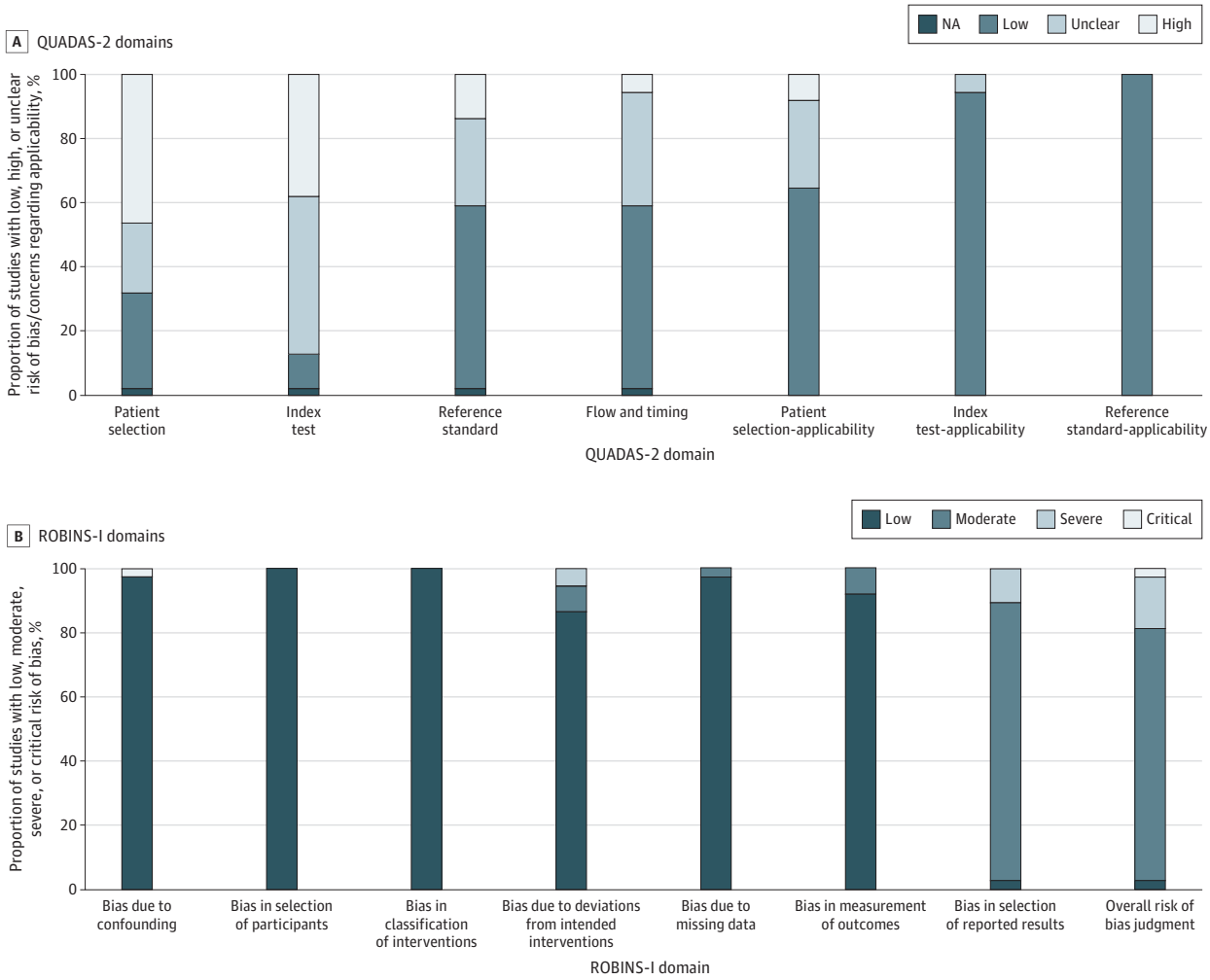
[b] Three studies reported on more than 1 CDSS or used the same CDSS in different modalities.

decision. The participants also considered it would be of small to moderate benefit if the CDSS would display more information on how it generates its decision and thought CDSSs could be of moderate to substantial benefit to support training and improve inexperienced clinicians' performance. Endo et al[32] invited study participants to give direct feedback on CDSS outputs by grading their relevance in the context of a specific task; 87% of the outputs were judged satisfactory. Additional human factor–related characteristics of the included studies can be found in eTable 11 in the Supplement.

Using QUADAS-2, 28 studies (76%) were rated as having high risk of bias in at least 1 of the 4 core domains, and none were considered to have a low risk of bias in all 4 core domains. Patient selection and the index test were the 2 domains most frequently found at high risk of bias. Using ROBINS-I, 6 studies (16%) were rated as having serious or critical risk of bias due to cofounding, deviation from the intended interventions, or likely selection of the reported results. Only 1 study was considered to be at low risk of bias in all 7 domains.[47] **Figure 2** shows the overall risk of bias assessment for each category of these tools.

Six studies (16%) reported private sector funding, and 12 (32%) gave no or unclear information about their source of funding. Only 2 studies (5%) referenced a study protocol.[47,54] The grey literature search retrieved 1 randomized clinical trial protocol (with expected completion after the

Figure 2. Distribution of the Risk of Bias Scores



The total of 100% represents 37 included studies in the Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) (A) and Risk of Bias in Non-Randomised Studies–Intervention (ROBINS-I) (B) domains.

present review's search period), 1 conference abstract (leading to a publication after the present review's search period),[56] and 1 conference abstract that did not lead to any publication.

Discussion

This systematic review found no robust evidence that the use of ML-based algorithms was associated with better clinician diagnostic performance. The evidence for any conclusion was weak because of a high risk of bias in many of the studies and a low number of study participants. Almost half of all results reported with statistical significance showed no significant difference in performance with or without the use of CDSSs. In studies conducted in a clearly reported representative clinical environment, this observation was even clearer, with 80% of the designated results showing no statistically significant change in performance. These findings corroborate the conclusions of several other studies assessing the outcome of CDSS use in mammogram screening across large populations, in which few or no benefits were found.[57-59] In a cross-specialty review like ours, expressing a straightforward judgment about the benefits of a CDSS is often difficult as it heavily depends on factors such as common clinical practice in a field or the prevalence of the target condition. This factor is the reason why we summarized the association between the use of CDSSs and clinician performance by metrics, as they enable readers to decide whether specific changes are desirable in their specialty. The interobserver agreement was the metric whose change appeared to be the most clearly associated with the use of CDSSs. The use of CDSSs also appeared to have a more marked association with increased performance for less experienced clinicians and to be associated with increased interobserver agreement between clinicians of different experience levels. In this way, CDSSs need not be solely used to outperform the most experienced clinicians but could be targeted by design toward those with less experience who may receive more benefit.

Little consideration was given to human factors in included studies. This outcome is surprising as human clinicians should be the main beneficiaries of the systems tested. In only 13 studies were the observers trained on the CDSS before the test. Given the likely existence of a learning (or trust) curve as observed by Rodríguez-Ruiz et al,[38] this omission might well have distorted some of the results. User feedback was reported in only 4 studies, hence hindering any iterative improvement in the human-computer interaction. This outcome is in contrast to other safety-critical industries, such as the aviation or energy sectors, where human factors principles have been commonly used for years.[60-64]

In all but 1 study in which the information was available, human operators decided to override at least some of the system recommendations, and it remains unclear to what extent human intelligence influences the overall system performance. These 2 observations highlight that computer simulations alone are insufficient to define the effectiveness and safety profile of a CDSS. In clinical situations in which humans have the responsibility for a diagnostic or therapeutic choice, they will, consciously or not, factor in other variables than the CDSS outputs and possibly prioritize their own clinical judgment in case of conflict. Therefore, it is the human processing of algorithm outputs, rather than the outputs themselves, that will affect patient care. Thus, it is important to evaluate this shared decision-making process rather than the CDSS stand-alone performance.

Many of the included studies were at high risk of bias, echoing the results of a recent review assessing studies comparing deep learning–based algorithms to clinicians.[4] This elevated risk of bias was mainly attributed to 3 factors: (1) the lack of prospectively or randomly selected case samples, (2) the absence during the test of clinical data otherwise available in real-life settings, and (3) the absence of a protocol. Moreover, the generalizability of the studies' findings was undermined by the absence of any power calculation and the median number of participants being only 4. In many cases, we also observed confusion between statistical significance at the patient and practitioner levels. Bootstrapping the clinical cases to produce a *P* value would not, for example, give any indication about the generalizability of findings to other clinicians. Instead, it would assess the likelihood that the same clinicians would display similar improvement with a new sample of patients.

In addition to the issues already outlined, there was a marked heterogeneity in the metrics used to assess CDSSs. Together, these inconsistencies make a reliable comparison of the different systems almost impossible. The issue of performance comparability is well known to the field and initiated the creation of data challenges, notably in medical image analysis, to evaluate how competing algorithms perform on common data sets.[65,66] This harmonization work should now be extended to the next phases of CDSS evaluation pathways, particularly when first used with human clinicians. Reporting guidelines would offer a practicable solution to this end.

Strengths and Limitations

The methodologic approach followed best practice standards for systematic reviews, and each step of the process was performed independently by at least 2 reviewers. This study is, to our knowledge, the first to put human clinicians, rather than the algorithms, at the forefront of a systematic review about the clinical use of ML-based CDSSs. This approach provides important information that nuances a commonly portrayed view that artificial intelligence may soon substantially improve clinician diagnostic performance across specialties. This approach also highlights the current lack of consideration of human factors when assessing the potential benefits of new CDSSs. In addition, this review provides material that can inform the development of further guidance on ML-based CDSS evaluation, complementing existing or upcoming reporting guidelines.[67-70] Such guidance will be particularly relevant for safety and effectiveness evaluations before the execution of large-scale clinical trials.

This review has limitations. It is possible that some relevant literature was not retrieved owing to (1) the heterogeneous description of the target CDSSs across medical specialties, (2) the use of commercial names only in many studies, and (3) the only recent categorization of this technology in specialized search engines (*machine learning* was added as a MeSH term in PubMed in 2016). We addressed these issues by conducting a forward and backward literature search of the included studies as well as an additional search for common or new commercial names. Given the broad range of CDSSs evaluated herein, certain inclusion criteria had to be defined very precisely, and some of these definitions are debatable because there is no broad consensus in the literature.

Conclusions

This systematic review of the literature provides findings to inform current and future debate about the evaluation of ML in health care. We found no robust evidence to suggest that the use of ML-based CDSSs is associated with improved diagnostic performance among clinicians in representative clinical environments. We also highlighted that most of the studies on this topic were at high or unclear risk of bias and had a low number of participants. In addition, we observed that the human operators almost always decided to override at least some of the CDSS recommendations. Therefore, we recommend more thorough evaluation of ML-based CDSSs and that more consideration be given to the human component of assisted diagnosis. These changes in practice should be guided by accepted principles of trial conduct and reporting to avoid repetition of errors noted in the current literature. Increased regulatory scrutiny also has an important role in ensuring a safe and efficient translation to the patient bedside. The results of this review should not be interpreted as tarnishing the prospects of ML-based diagnostic CDSSs. Rather, we encourage qualitative improvements in future research. Better methodologies and evaluations would allow CDSSs to showcase their full potential and ultimately improve patient care.

ARTICLE INFORMATION

Accepted for Publication: January 20, 2021.

Published: March 11, 2021. doi:10.1001/jamanetworkopen.2021.1276

Corresponding Author: Baptiste Vasey, MMed, Nuffield Department of Surgical Sciences, University of Oxford, Headington, Oxford OX3 9DU, United Kingdom (baptiste.vasey@nds.ox.ac.uk).

Author Affiliations: Nuffield Department of Surgical Sciences, University of Oxford, Oxford, United Kingdom (Vasey, Taylor, Marlow, McCulloch); Department of Radiology, University of Cambridge, Cambridge, United Kingdom (Ursprung); Faculty of Medicine, Imperial College London, London, United Kingdom (Beddoe); Oxford University Hospitals NHS Foundation Trust, Oxford, United Kingdom (Marlow); Department of Surgery, Maimonides Medical Center, Brooklyn, New York (Bilbro); Critical Care Research Group, Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, United Kingdom (Watkinson).

Author Contributions: Mr Vasey had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. Dr McCulloch is the study's guarantor.

Concept and design: Vasey, Ursprung, Marlow, Bilbro, McCulloch.

Acquisition, analysis, or interpretation of data: Vasey, Ursprung, Beddoe, Taylor, Marlow, Bilbro, Watkinson.

Drafting of the manuscript: Vasey, Marlow.

Critical revision of the manuscript for important intellectual content: Ursprung, Beddoe, Taylor, Marlow, Bilbro, Watkinson, McCulloch.

Statistical analysis: Vasey.

Administrative, technical, or material support: Vasey, Beddoe, Taylor.

Supervision: Watkinson, McCulloch.

Conflict of Interest Disclosures: Mr Vasey reported participation in a CS Digital Health Equity Fund (participations sold in January 2020) outside the submitted work. Mr Ursprung reported a scholarship from Cambridge Commonwealth, European & International Trust Scholarship during the conduct of the study. Dr Watkinson reported receiving grants from the National Institute for Health Research (NIHR) during the conduct of the study; grants from the NIHR, Wellcome, and Sensyne Health; and personal fees from Sensyne Health. He was chief medical officer for Sensyne Health and holds shares in the company outside the submitted work. No other disclosures were reported.

Funding/Support: Mr Vasey is supported by the Berrow Foundation (Lincoln College, University of Oxford); Mr Ursprung is supported by the Cambridge Commonwealth, European & International Trust; and Dr Watkinson is supported by the NIHR Biomedical Research Centre, Oxford.

Role of the Funder/Sponsor: The funding sources played no role in the study design, data collection, analysis, or decision to submit for publication.

Additional Contributions: Tatjana Petrinic, outreach librarian (Bodleian Libraries, University of Oxford), helped in designing the search strategy and providing guidance throughout all stages of the review; no financial compensation outside of salary was provided. Mr Vasey thanks the Berrow Foundation, Lincoln College, University of Oxford, for their support, without which this research would not have been possible.

REFERENCES

1. CBInsights. State of healthcare report Q2'20: sector and investment trends to watch. Accessed January 24, 2021. https://www.cbinsights.com/research/report/healthcare-trends-q2-2020/

2. American College of Radiology Data Science Institute. FDA cleared AI algorithms. Accessed September 10, 2020. https://www.acrdsi.org/DSI-Services/FDA-Cleared-AI-Algorithms

3. Liu X, Faes L, Kale AU, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit Health*. 2019;1(6):e271-e297. doi:10.1016/S2589-7500(19)30123-2

4. Nagendran M, Chen Y, Lovejoy CA, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ*. 2020;368:m689. "https://www.bmj.com/content/368/bmj.m689" doi:10.1136/bmj.m689

5. Haselton MG, Nettle D, Murray DR. The evolution of cognitive bias. In: Buss DM, ed. *The Handbook of Evolutionary Psychology*. John Wiley & Sons Inc; 2015:968-987. doi:10.1002/9781119125563.evpsych241

6. Croskerry P. The importance of cognitive errors in diagnosis and strategies to minimize them. *Acad Med*. 2003;78(8):775-780. doi:10.1097/00001888-200308000-00003

7. Cook RI, Woods DD. Operating at the sharp end: the complexity of human error. In: Bogner MS, ed. *Human Error in Medicine*. CRC Press; 2018.

8. Riches N, Panagioti M, Alam R, et al. The effectiveness of electronic differential diagnoses (DDX) generators: a systematic review and meta-analysis. *PLoS One*. 2016;11(3):e0148991. doi:10.1371/journal.pone.0148991

9. Bright TJ, Wong A, Dhurjati R, et al. Effect of clinical decision-support systems: a systematic review. *Ann Intern Med*. 2012;157(1):29-43. doi:10.7326/0003-4819-157-1-201207030-00450

10. Jaspers MWM, Smeulers M, Vermeulen H, Peute LW. Effects of clinical decision-support systems on practitioner performance and patient outcomes: a synthesis of high-quality systematic review findings. *J Am Med Inform Assoc*. 2011;18(3):327-334. doi:10.1136/amiajnl-2011-000094

11. Garg AX, Adhikari NKJ, McDonald H, et al. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. *JAMA*. 2005;293(10):1223-1238. doi:10.1001/jama.293.10.1223

12. Varghese J, Kleine M, Gessner SI, Sandmann S, Dugas M. Effects of computerized decision support system implementations on patient outcomes in inpatient care: a systematic review. *J Am Med Inform Assoc*. 2018;25(5):593-602. doi:10.1093/jamia/ocx100

13. Samuel AL. Some studies in machine learning using the game of checkers. *IBM J Res Dev*. 1959;3(3):210–29. doi:10.1147/rd.33.0210

14. Moher D, Liberati A, Tetzlaff J, Altman DG; PRISMA Group. Preferred Reporting Items for Systematic Reviews and Meta-analyses: the PRISMA statement. *BMJ*. 2009;339:b2535. doi:10.1136/bmj.b2535

15. Covidence systematic review software, Veritas Health Innovation, Melbourne, Australia. Accessed January 24, 2021. http://www.covidence.org

16. Whiting PF, Rutjes AWS, Westwood ME, et al; QUADAS-2 Group. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med*. 2011;155(8):529-536. doi:10.7326/0003-4819-155-8-201110180-00009

17. Sterne JAC, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ*. 2016;355:i4919. "https://www.bmj.com/content/355/bmj.i4919" doi:10.1136/bmj.i4919

18. Aissa J, Schaarschmidt BM, Below J, et al. Performance and clinical impact of machine learning based lung nodule detection using vessel suppression in melanoma patients. *Clin Imaging*. 2018;52:328-333. doi:10.1016/j.clinimag.2018.09.001

19. Aslantas A, Dandil E, Sağlam S, Çakiroğlu M. CADBOSS: a computer-aided diagnosis system for whole-body bone scintigraphy scans. *J Cancer Res Ther*. 2016;12(2):787-792. doi:10.4103/0973-1482.150422

20. Bargalló X, Santamaría G, Del Amo M, et al. Single reading with computer-aided detection performed by selected radiologists in a breast cancer screening program. *Eur J Radiol*. 2014;83(11):2019-2023. doi:10.1016/j.ejrad.2014.08.010

21. Barinov L, Jairaj A, Becker M, et al. Impact of data presentation on physician performance utilizing artificial intelligence-based computer-aided diagnosis and decision support systems. *J Digit Imaging*. 2019;32(3):408-416. doi:10.1007/s10278-018-0132-5

22. Bartolotta TV, Orlando A, Cantisani V, et al. Focal breast lesion characterization according to the BI-RADS US lexicon: role of a computer-aided decision-making support. *Radiol Med*. 2018;123(7):498-506. doi:10.1007/s11547-018-0874-7

23. Bien N, Rajpurkar P, Ball RL, et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of MRNet. *PLoS Med*. 2018;15(11):e1002699. doi:10.1371/journal.pmed.1002699

24. van den Biggelaar FJHM, Kessels AGH, van Engelshoven JMA, Boetes C, Flobbe K. Computer-aided detection in full-field digital mammography in a clinical population: performance of radiologist and technologists. *Breast Cancer Res Treat*. 2010;120(2):499-506. doi:10.1007/s10549-009-0409-y

25. Blackmon KN, Florin C, Bogoni L, et al. Computer-aided detection of pulmonary embolism at CT pulmonary angiography: can it improve performance of inexperienced readers? *Eur Radiol*. 2011;21(6):1214-1223. doi:10.1007/s00330-010-2050-x

26. Cha KH, Hadjiiski LM, Cohan RH, et al. Diagnostic accuracy of CT for prediction of bladder cancer treatment response with and without computerized decision support. *Acad Radiol*. 2019;26(9):1137-1145. doi:10.1016/j.acra.2018.10.010

27. Chabi ML, Borget I, Ardiles R, et al. Evaluation of the accuracy of a computer-aided diagnosis (CAD) system in breast ultrasound according to the radiologist's experience. *Acad Radiol*. 2012;19(3):311-319. doi:10.1016/j.acra.2011.10.023

28. Cho E, Kim EK, Song MK, Yoon JH. Application of computer-aided diagnosis on breast ultrasonography: evaluation of diagnostic performances and agreement of radiologists according to different levels of experience. *J Ultrasound Med*. 2018;37(1):209-216. doi:10.1002/jum.14332

29. Choi JH, Kang BJ, Baek JE, Lee HS, Kim SH. Application of computer-aided diagnosis in breast ultrasound interpretation: improvements in diagnostic performance according to reader experience. *Ultrasonography*. 2018; 37(3):217-225. doi:10.14366/usg.17046

30. Choi JS, Han BK, Ko ES, et al. Effect of a deep learning framework–based computer-aided diagnosis system on the diagnostic performance of radiologists in differentiating between malignant and benign masses on breast ultrasonography. *Korean J Radiol*. 2019;20(5):749-758. doi:10.3348/kjr.2018.0530

31. Cole EB, Zhang Z, Marques HS, Edward Hendrick R, Yaffe MJ, Pisano ED. Impact of computer-aided detection systems on radiologist accuracy with digital mammography. *AJR Am J Roentgenol*. 2014;203(4):909-916. doi:10.2214/AJR.12.10187

32. Endo M, Aramaki T, Asakura K, et al. Content-based image-retrieval system in chest computed tomography for a solitary pulmonary nodule: method and preliminary experiments. *Int J Comput Assist Radiol Surg*. 2012;7(2): 331-338. doi:10.1007/s11548-011-0668-z

33. Engelke C, Schmidt S, Auer F, Rummeny EJ, Marten K. Does computer-assisted detection of pulmonary emboli enhance severity assessment and risk stratification in acute pulmonary embolism? *Clin Radiol*. 2010;65(2): 137-144. doi:10.1016/j.crad.2009.10.007

34. Giannini V, Mazzetti S, Armando E, et al. Multiparametric magnetic resonance imaging of the prostate with computer-aided detection: experienced observer performance study. *Eur Radiol*. 2017;27(10):4200-4208. doi:10.1007/s00330-017-4805-0

35. Hwang EJ, Park S, Jin KN, et al; DLAD Development and Evaluation Group. Development and validation of a deep learning–based automated detection algorithm for major thoracic diseases on chest radiographs. *JAMA Netw Open*. 2019;2(3):e191095. doi:10.1001/jamanetworkopen.2019.1095

36. Lindsey R, Daluiski A, Chopra S, et al. Deep neural network improves fracture detection by clinicians. *Proc Natl Acad Sci U S A*. 2018;115(45):11591-11596. doi:10.1073/pnas.1806905115

37. Park HJ, Kim SM, La Yun B, et al. A computer-aided diagnosis system using artificial intelligence for the diagnosis and characterization of breast masses on ultrasound: added value for the inexperienced breast radiologist. *Medicine (Baltimore)*. 2019;98(3):e14146. doi:10.1097/MD.0000000000014146

38. Rodríguez-Ruiz A, Krupinski E, Mordang JJ, et al. Detection of breast cancer with mammography: effect of an artificial intelligence support system. *Radiology*. 2019;290(2):305-314. doi:10.1148/radiol.2018181371

39. Romero C, Varela C, Muñoz E, Almenar A, Pinto JM, Botella M. Impact on breast cancer diagnosis in a multidisciplinary unit after the incorporation of mammography digitalization and computer-aided detection systems. *AJR Am J Roentgenol*. 2011;197(6):1492-1497. doi:10.2214/AJR.09.3408

40. Samulski M, Hupse R, Boetes C, Mus RDM, den Heeten GJ, Karssemeijer N. Using computer-aided detection in mammography as a decision support. *Eur Radiol*. 2010;20(10):2323-2330. doi:10.1007/s00330-010-1821-8

41. Sanchez Gómez S, Torres Tabanera M, Vega Bolivar A, et al. Impact of a CAD system in a screen-film mammography screening program: a prospective study. *Eur J Radiol*. 2011;80(3):e317-e321. doi:10.1016/j.ejrad.2010.08.031

42. Sayres R, Taly A, Rahimy E, et al. Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy. *Ophthalmology*. 2019;126(4):552-564. doi:10.1016/j.ophtha.2018.11.016

43. Shimauchi A, Giger ML, Bhooshan N, et al. Evaluation of clinical breast MR imaging performed with prototype computer-aided diagnosis breast MR imaging workstation: reader study. *Radiology*. 2011;258(3):696-704. doi:10.1148/radiol.10100409

44. Sohns C, Angic BC, Sossalla S, Konietschke F, Obenauer S. CAD in full-field digital mammography-influence of reader experience and application of CAD on interpretation of time. *Clin Imaging*. 2010;34(6):418-424. doi:10.1016/j.clinimag.2009.10.039

45. Steiner DF, MacDonald R, Liu Y, et al. Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer. *Am J Surg Pathol*. 2018;42(12):1636-1646. doi:10.1097/PAS.0000000000001151

46. Stoffel E, Becker AS, Wurnig MC, et al. Distinction between phyllodes tumor and fibroadenoma in breast ultrasound using deep learning image analysis. *Eur J Radiol Open*. 2018;5:165-170. doi:10.1016/j.ejro.2018.09.002

47. Sun L, Li Y, Zhang YT, et al. A computer-aided diagnostic algorithm improves the accuracy of transesophageal echocardiography for left atrial thrombi: a single-center prospective study. *J Ultrasound Med*. 2014;33(1):83-91. doi:10.7863/ultra.33.1.83

48. Sunwoo L, Kim YJ, Choi SH, et al. Computer-aided detection of brain metastasis on 3D MR imaging: observer performance study. *PLoS One*. 2017;12(6):e0178265. doi:10.1371/journal.pone.0178265

49. Tang FH, Ng DKS, Chow DHK. An image feature approach for computer-aided detection of ischemic stroke. *Comput Biol Med*. 2011;41(7):529-536. doi:10.1016/j.compbiomed.2011.05.001

50. Taylor JC, Romanowski C, Lorenz E, Lo C, Bandmann O, Fenner J. Computer-aided diagnosis for (^{123}I)FP-CIT imaging: impact on clinical reporting. *EJNMMI Res*. 2018;8(1):36. doi:10.1186/s13550-018-0393-5

51. Vassallo L, Traverso A, Agnello M, et al. A cloud-based computer-aided detection system improves identification of lung nodules on computed tomography scans of patients with extra-thoracic malignancies. *Eur Radiol*. 2019;29(1):144-152. doi:10.1007/s00330-018-5528-6

52. Watanabe AT, Lim V, Vu HX, et al. Improved cancer detection using artificial intelligence: a retrospective evaluation of missed cancers on mammography. *J Digit Imaging*. 2019;32(4):625-637. doi:10.1007/s10278-019-00192-5

53. Way T, Chan HP, Hadjiiski L, et al. Computer-aided diagnosis of lung nodules on CT scans: ROC study of its effect on radiologists' performance. *Acad Radiol*. 2010;17(3):323-332. doi:10.1016/j.acra.2009.10.016

54. Zhang J, Wang Y, Yu B, Shi X, Zhang Y. Application of computer-aided diagnosis to the sonographic evaluation of cervical lymph nodes. *Ultrason Imaging*. 2016;38(2):159-171. doi:10.1177/0161734615589080

55. IMDRF Software as Medical Device (SaMD) Working Group. "Software as a medical device:" possible framework for risk categorization and corresponding considerations. Published September 18, 2014. Accessed January 24, 2021. https://www.fdanews.com/ext/resources/files/10-14/10-14-IMDRF-SaMD.pdf?1520753258

56. Han SS, Park I, Eun Chang S, et al. Augmented intelligence dermatology: deep neural networks empower medical professionals in diagnosing skin cancer and predicting treatment options for 134 skin disorders. *J Invest Dermatol*. 2020;140(9):1753-1761. doi:10.1016/j.jid.2020.01.019

57. Lehman CD, Wellman RD, Buist DSM, Kerlikowske K, Tosteson AN, Miglioretti DL; Breast Cancer Surveillance Consortium. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA Intern Med*. 2015;175(11):1828-1837. doi:10.1001/jamainternmed.2015.5231

58. Fenton JJ, Abraham L, Taplin SH, et al; Breast Cancer Surveillance Consortium. Effectiveness of computer-aided detection in community mammography practice. *J Natl Cancer Inst*. 2011;103(15):1152-1161. doi:10.1093/jnci/djr206

59. Fenton JJ, Xing G, Elmore JG, et al. Short-term outcomes of screening mammography using computer-aided detection: a population-based study of Medicare enrollees. *Ann Intern Med*. 2013;158(8):580-587. doi:10.7326/0003-4819-158-8-201304160-00002

60. Harris D, Stanton NA, Marshall A, et al. Using SHERPA to predict design-induced error on the flight deck. *Aerosp Sci Technol*. 2005;9(6):525-532. doi:10.1016/j.ast.2005.04.002

61. Isaac A, Shorrock ST, Kirwan B. Human error in European air traffic management: the HERA project. *Reliab Eng Syst Saf*. 2002;75(2):257-272. doi:10.1016/S0951-8320(01)00099-0

62. Thomas LC, Rantanen EM. Human factors issues in implementation of advanced aviation technologies: a case of false alerts and cockpit displays of traffic information. *Theor Issues Ergon Sci*. 2006;7(5):501-523. doi:10.1080/14639220500090083

63. Stanton NA, Salmon P, Jenkins D, Walker G. *Human Factors in the Design and Evaluation of Central Control Room Operations*. CRC Press; 2009. doi:10.1201/9781439809921

64. Carvalho PVR, dos Santos IL, Gomes JO, Borges MRS, Guerlain S. Human factors approach for evaluation and redesign of human–system interfaces of a nuclear power plant simulator. *Displays*. 2008;29(3):273-284. doi:10.1016/j.displa.2007.08.010

65. Price K. Anything you can do, I can do better (no you can't)…. *Comput Vis Graph Image Process*. 1986;36(2):387-391. doi:10.1016/0734-189X(86)90083-6

66. van Ginneken B, Kerkstra S, Meakin J. Challenges: Grand Challenge. Accessed September 10, 2020. https://grand-challenge.org/challenges/

67. Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK; SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI Extension. *BMJ*. 2020;370:m3164. doi:10.1136/bmj.m3164

68. Rivera SC, Liu X, Chan A-W, Denniston AK, Calvert MJ; SPIRIT-AI and CONSORT-AI Working Group. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI Extension. *BMJ*. 2020;370:m3210. doi:10.1136/bmj.m3210

69. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med*. 2015;162(1):55-63. doi:10.7326/M14-0697

70. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. *Lancet*. 2019;393(10181):1577-1579. doi:10.1016/S0140-6736(19)30037-6

SUPPLEMENT.

eAppendix 1. Systematic Review Protocol
eAppendix 2. Modified Search Strategies
eTable 1. Metrics Used to Evaluate the Impact of ML-Based CDSS on Human Performance
eTable 2. Impact of ML-Based CDSS on Clinician Performance in Patients or Lesions Subgroup
eTable 3. Complete List of the Included Studies' Results for the Primary Outcome
eTable 4. Impact on Clinician Performance of the Six ML-Based CDSS Evaluated in Representative Clinical Environment
eTable 5. Association Between Clinicians' Level of Experience and Performance Changes When Using ML-Based CDSS
eTable 6. Impact on Clinician Performance of ML-Based CDSS According to the Reader Paradigm (First Reader/Second Reader)
eTable 7. Impact on Clinician Performance of ML-Based CDSS According to the Mathematical Model Used (Neural Networks/Other Models)
eTable 8. Impact on Clinician Performance of ML-Based CDSS According to the Outputs' Level of Support (Single Output/Explanatory Output)
eTable 9. Impact of the Human Contribution on the System Performance in Patients or Lesions Subgroups
eTable 10. Complete List of the Included Studies' Results for the Secondary Outcome (Assisted Human Performance vs Stand-Alone Computer Performance)
eTable 11. Characteristics Relevant to the Human Factors Evaluation of the Included Studies